Neurocomputing 279 (2018) 54–62

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Effective community division based on improved spectral clustering

Yi Xu [a], Zhi Zhuang [a], Weimin Li [a,*], Xiaokang Zhou [b]

[a] School of Computer Engineering and Technology, Shanghai University, Shanghai, China
[b] Faculty of Data Science, Shiga University, Hikone, Japan

ARTICLE INFO

Article history:
Received 6 October 2016
Revised 2 March 2017
Accepted 18 June 2017
Available online 21 November 2017

Keywords:
Spectral clustering
Attribute and relationship
Community division
Particle swarm optimization (PSO)
Simulated Annealing (SA)

ABSTRACT

Not only does attribute of nodes affect the effectiveness and efficiency of community division, but also the relationship of them has a great impact on it. Clusters of arbitrary shape can be identified by the Spectral Clustering (SC). However, k-means clustering used in SC still could result in local optima, and the parameters in Radial Basis Function need to be determined by trial and error. In order to make such algorithm better fit into community division of social network, we try to merge attribute and relationship of node and optimize the ability of spectral clustering to get the global solution, thus a new community clustering algorithm called Spectral Clustering Based on Simulated Annealing and Particle swarm optimization (SCBSP) is proposed. The proposed algorithm is adapted to social networking division. In related experiments, the proposed algorithm, which enhances the global searching ability, has better global convergence and makes better performance in community division than original spectral clustering.

1. Introduction

Recently, with the rapid development of Internet technology, big data generated by social networks such as Twitter and Facebook are urgently needed to be explored and analyzed. Due to the large scale of social networks, it is important to turn them into smaller ones through community division methods. Thus, the effectiveness and efficiency of community division has become an important issue in this regard [1]. Dividing social network by employing effective clustering algorithms can have a wide range of application in our real-life world. Social networks have underlying clusters according to individuals' attribute and relationship. Usually, people in a certain cluster have similar interests and personality. Therefore, the common interest in a certain cluster can be deduced by analyzing a small sample in it, and then companies can recommend their products with higher accuracy and lower risk. In the field of education, educators can adapt their educational method to different level of students classified by such algorithms. In criminal psychology, if there are several criminals in a cluster, police can pay more attention to other individuals in this cluster, who may also have high potential of committing a crime. What's more, an effective clustering algorithm can be applied to other fields such as image processing and computer vision.

There are many kinds of clustering algorithms used to detect community. Classical algorithms are popular. For example, k-means, k-medoids and K-Harmonic [2] are based on the node attributes, while Fast-Newman algorithm [3] was proposed on the basis of the modularity focuses on relationship. Hierarchical clustering is another kind of algorithm to divide the community, such as Louvain algorithm for large networks [4], Girvan Newman algorithm [5] and link community [6]. Scott and Smyth combined both Q modularity and spectral clustering together through Q-Laplacian matrix and local greedy heuristic search [7]. Shihua Zhang raised a new modularity function based on generalizing Q modularity and fuzzy c-means clustering [8]. In addition, there are some other community clustering algorithms: Density-based spatial clustering which has a wide range of application in urban planning and marketing [9], graph-skeleton based clustering [10] and Structural Clustering Algorithm for Networks (SCAN) based on the edge density and Clustering Centrality [11].

Most of the clustering algorithms separate the attribute and relationship of nodes while clustering a complex graph. But in fact, they both affect the result of community division. For example, in co-authorship network, author's researching direction and current partnership both affect the frequency and probability of cooperation. And in social network, similar interest and friendship make two users close to each other. Therefore, more exact information can be obtained by taking attribute and relationship of node into account together. To some degree, it can be important for social network research. Actually, there have been some studies in this field [12]. On the basis of these factors, we aimed to improve a clustering algorithm adapted to this kind of community network. Thus, spectral clustering comes into our mind.

* Corresponding author.
E-mail addresses: wmli@shu.edu.cn (W. Li), zhou@biwako.shiga-u.ac.jp (X. Zhou).

https://doi.org/10.1016/j.neucom.2017.06.085
0925-2312/

In this paper, we aim at improving Spectral Clustering (SC) to increase its efficiency in community division of social network. Intensified research on spectral clustering leads to explosive development and improvement over the past several years [13], which is easy to implement and reasonably fast. However, traditional SC is sensitive to the initial data. What's more, after feature decomposition, traditional SC chooses k-means clustering to cluster with the eigenvectors-matrix, while k-means clustering is easy to converge to a local optimal solution. Taking these problems above into account, and given the rapid convergence of Particle Swarm optimization (PSO) and good ability to search the global optimal solution of Simulated Annealing (SA), a new algorithm called Spectral Clustering Based on the Simulated annealing and Particle swarm optimization (SCBSP) was proposed. Actually, the proposed algorithm is adapted to the community division of social network. From experiments which are going to be explained in detail later, SCBSP really do make a step forward in community division of real-life social networks such as Sina Microblog and Facebook.

In our study, we first improve the traditional spectral clustering itself. We implement traditional spectral clustering and apply it to cluster simple real-life social network, in which the performance of traditional SC is not very satisfying. To some degree it is because k-means algorithm is easy to reach a local optima. One of the good replacements of k-means is PSO, which has high convergence speed and good performance in low-dimensional vector space, combined with SA, which has the ability of finding global optima. SCBSP proposed in this paper is based on merging these two algorithms with SC.

Next, we revise the step of preprocessing data in order to make our improved SC better fit into community division. When doing community division, it is important to take both attribute of nodes and relationship between them into consideration. However, traditional similarity matrix just cares about the differences between the attributes of nodes, and it ignores the relationship between nodes in the situation of community division. Thus we define a new similarity matrix which merges attribute and relationship.

Finally, we conduct several experiments on both randomly generated data and real-life social network to evaluate our method. The experiment results show that the method has both high efficiency and high accuracy.

The rest of this paper is organized as follows: Section 2 introduces related work; Section 3 explains our proposed algorithm in detail which begins with a brief review on SC, PSO and SA; The experimental results on random generated data and real-life data from famous social networks are presented in Section 4, and Section 5 concludes this paper.

2. Related work

Social networks are ubiquitous, and researchers have investigated a growing number of data generated by social networks. Yet, most existing measuring methods do not take both attribute and relationship of nodes into consideration, thus they do not fully capture the richness of the information contained in the data [14,15]. Most methods focus on improving the k-means, k-medoids, Newman and Girvan, Density-based methods and so on.

The classical partitioning methods for clustering are k-means and k-medoids, they are easy to implement but are based on complex mathematical theory. These classical algorithms are foundation of many other clustering algorithms. In k-means algorithm, clusters are represented by a mean value and object exchanging stops if the average distance from objects to their cluster's mean value converges to a minimum value [16]. K-medoids algorithm represents each cluster by an actual object in it. However, as is known to all, the original k-means proposed by James MacQueen is easy to converge to a local optimal solution and sensitive to the initial data [14]. In k-harmonic means, harmonic means function which applies distance from the data point to all clustering centers is used to solve the problem that clustering result is sensitive to initial value instead of the minimum distance [2]. Although the problem about initial data is solved, another problem still exists. On the basis, the k-harmonic means was improved by the Simulated Annealing called K-Harmonic Means Clustering with Simulated Annealing [17]. In the K-Harmonic Means Clustering, simulated annealing is used to search the global solution whenever a new result is obtained by k-harmonic means.

In the classical algorithms, the number of clusters must be selected by researchers and is usually tried for times by tests. So, researchers sought for methods which can automatically choose number of clusters. Unlike the k-means that needs to know the number of clusters first of all, an objective function for graph clustering was raised by Newman and Girvan called Q function. In this method, the number of clusters can be automatic selected, which avoid trying cluster numbers for several times before getting a better result. In other word, Q function has higher value when combined with good clustering method. While because of the high complexity of Girvan-Newman algorithm, Newman proposed another algorithm based on the Q function called Fast-Newman algorithm [3,18]. What's more, because spectral clustering is popular for its process of recursively splitting the graph, Scott et al combined both Q function and spectral clustering together. From the experimental results, the two novel algorithms proposed by Scott are efficient and effective [7]. The first algorithm directly searches for global maximum of Q by performing eigenvector decomposition on a matrix called Q-Laplacian matrix, while Newman's algorithm improves the maximum of Q by local iteration. The second algorithm uses a local greedy heuristic search which is similar to Newman's method. At the same time, for the popularity of spectral clustering, a lot of researcher has explored the algorithm, such as the Shi and Malik algorithm, the Kannan, Vempala and Vetta algorithm, the Ng, Jordan and Weiss algorithm [19,20,14] and other more efforts [21–24]. In addition, the ideal of Q function is also applied to identification of overlapping community structure. Shihua Zhang combined a new modularity function based on generalizing Q function and fuzzy c-means clustering [8].

Other researchers sought for methods different from the classical algorithms, they proposed algorithms such as density-based algorithms based on the structure of data. To some degree, density-based methods are the same as Girvan-Newman methods, users don't need to know the number of clusters at the beginning. Density-based methods are more sensitive to initial parameters than Girvan-Newman methods, but they can identify the noise in the data. When several objects are close to each other, they form dense clusters, and they are separated from each other by regions with low density of objects. Thus dense clusters can be detected by finding such low-density regions. DBSCAN (Density-based spatial clustering of application with noise) is the most representative method in this field. In addition, density-based methods can also be applied to spatial clustering, such as clustering in Geo-Social Networks [16]. Because most traditional density-based methods have difficulty in detecting communities of arbitrary and sometimes extreme shape, for example evenly distributed nodes, and they usually have difficulty identifying central nodes and outlying ones, gSkeletonClu (graph-skeleton based clustering) was proposed for community division [10]. The gSkeletonClu algorithm projects the network to its Core-Connected Maximal Spanning Tree (CCMST) and finds its core nodes to cluster this network. This novel algorithm can also avoid the problem of the chaining effect and resolution limit faced by typical MST-based clustering algorithms and modularity-based algorithms.

Recently, much ink has been spilled onto new methods like spectral clustering [13]. These methods focus on improving the

traditional spectral clustering. However, although spectral clustering has been applied to cluster social network [25], the method of spectral clustering merging attribute and relationship of node has not been employed before. Spectral clustering has a unique preprocess step, which we want to take advantage of to merge attribute and relationship. Thus we choose to improve SC to make it a more powerful method and better fit into community division. In this paper, the proposed algorithm is different from the algorithms mentioned above, and the experimental results show that the proposed algorithm has a good result.

3. Spectral clustering based on simulated annealing and particle swarm optimization

3.1. Spectral clustering

The method of spectral clustering originated from the spectral graph division, and the essential part of spectral clustering is to construct and make use of the Laplace matrix to reduce dimensionality of dataset. To some degree, this step is equivalent to the reconstruction of the original sample space.

In the spectral clustering algorithm, clustering problem can be regarded as partitioning an undirected graph. Suppose an undirected graph as *G (V, E), V* represents the nodes in the graph *G*, and *E* is the relationship between nodes. Usually, *E* is represented by an adjacent matrix. What's more, the criteria of classification has various ways, such as Normalized cut, Ratio cut, Average cut and so on, but the goal of the criteria is to make the correlation of two nodes larger in the same subgraph, and smaller in different subgraphs [26].

In the step of preprocessing data, the graph should be transferred into similarity matrix and further transferred into Laplace matrix. Similarity matrix can be referred to as affinity matrix, which is usually represented by *S*, the element in *i*th row and *j*th column of the similarity matrix is defined as:

$$S_{ij} = e^{-\left(\frac{d(v_i,v_j)}{2\sigma^2}\right)} \tag{1}$$

where v_i represents the *i*th node, $d(v_i,v_j)$ is the distance between node v_i and v_j, usually uses Euclidean distance to represent, σ is the scale parameter for the zoom level of the distance between two nodes [26,27].

Based on the statement above, the basic process of spectral clustering can be expressed as follows:

In this paper, the goal of the algorithm is to find a community division represented by $C=\{C_1,C_2,\ldots C_k\}$, and make the sum of differences smaller within the cluster. The sum of difference within the cluster is as follows:

$$J_c = \sum_{j=1}^{k} \sum_{i \in C_j} (R_i - CX_j)^2 \tag{4}$$

where the CX_j is a *k*-dimensional vector, $CX_j=(CX_{j1}, CX_{j2}, \ldots, CX_{jk})$ is on behalf of each center of cluster, and satisfies the Eq. (5).

$$CX_j = \frac{1}{n_j} \sum_{i \in C_j} R_i \tag{5}$$

where n_j is the number of nodes in *j*th cluster.

3.2. Improvement on spectral clustering in community division

The relationship of nodes is often overlooked in the typical spectral clustering algorithms and the contrast between the nodes ought to be concerned about when evaluating similarity matrix. Therefore, the goal is establishing the cluster model of the graph containing both attribute of nodes and relationship between them. The construction of the similarity matrix is different from the original spectral clustering.

The original spectral clustering calculates the similarity matrix by Radial Basis Function [26]. However, taking the uncertainty of the parameters σ in Radial Basis Function into consideration, an acceptable solution needs to be tried by tests. Thus it is inconvenient to use Radial Basis Function Though Hongjie Jia et al. proposed the Self-Turning algorithm in which parameters in similarity matrix can be determined automatically based on shared nearest neighborhood [28], Laviers et al. also proposed an approach to determine the parameter σ dynamically [29], and they still depend on this parameter. To solve the above problem, we proposed a method which does not use the parameter σ or Radial Basis Function.

Meanwhile, combined with PSO which has high speed of convergence and SA which has superior ability of obtaining global optima, a new clustering algorithm called Spectral Clustering Based on Simulated Annealing and Particle Swarm Optimization (SCBSP) is proposed.

3.3. A review of particle swarm optimization and simulated annealing

In the PSO algorithm, it is assumed that our problem is in a *k*-dimensional search space. The particle group consists of *n* particles, and each particle's position represents an existing solution in the *k*-dimensional search space. The position of *i*th particle is a *k*-dimensional vector $X_i=\{X_{i1}, X_{i2} \ldots X_{ik}\}$ and the moving speed of *i*th particle is another *k*-dimensional vector $V_i=\{V_{i1}, V_{i2}, \ldots, V_{ik}\}$. The individual historical optimal value of the *i*th particle is $P_i=\{P_{i1}, P_{i2}, \ldots, P_{ik}\}$ and the global optimal value can be defined as a *k*-dimensional vector $P_g=\{P_{g1}, P_{g2}, \ldots, P_{gk}\}$. Thus, each particle can update their speed and position according to the following formula:

$${V_{ij}}^{t+1} = \omega {V_{ij}}^{t} + \eta_1 r_1 \left(P_{ij} - {X_{ij}}^{t}\right) + \eta_2 r_2 \left(P_{gj} - {X_{ij}}^{t}\right) \tag{6}$$

$${X_{ij}}^{t+1} = {X_{ij}}^{t} + {V_{ij}}^{t+1} \tag{7}$$

where ω is inertia factor, $\varpi \in [0.1, 0.9]$. $\eta 1$, $\eta 2 > 0$ and are known as acceleration factor, usually $\eta 1 = \eta 2 = 2$, $r1$ and $r2$ is random number in $[0, 1]$. In the process of iteration, speed and position of particle can be confined within a certain range. In each step of iteration, both individual optimal solution and global optimal solution are updated. At the end of the iteration, P_g is the answer of global optimal solution.

Although PSO has fast convergence rate, the result of population evolution relies on the global extremum and individual extremum. To some degree, it is inevitable that PSO falls into local optimum easily. To solve this problem, Simulated annealing (SA) is introduced to assist Particle swarm optimization. SA is a stochastic optimization algorithm based on Monte Carlo iterative strategy. The algorithm is the simulation of physical annealing process, which allows the adjustment of the searching direction under the control of a decreasing temperature. Based on Metropolis guidelines, SA may accept a worse state at a certain probability in some steps while going towards better state on a global scale. Finally SA can obtain the global optimal solution [30]. To some degree, SA is an algorithm with low efficiency. However, a more accurate and efficient strategy called SP which combines SA and PSO can be applied to the SCBSP after dimensionality reduction. The description of SP will be introduced in detail in the step of SCBSP.

3.4. The main process of SCBSP

The first step is improving the similarity matrix. Suppose that a social network is a graph $G=(V, E)$, where *V* represents nodes in

the graph, and the relationship between nodes is E. Assume that there are n nodes, and each node has m-dimensional properties, so each node consists of an m-dimensional vector $A(V_i)=(a_{i1}, a_{i2}, \ldots, a_{im})$. Then, the properties of all nodes can be represented by an n^*m matrix called A, where A_{ij} denotes the jth property of ith node. And all the relationship can be denoted by an n^*n matrix W. Then, for convenience, Euclidean distance is employed here to represent the discrepancy between two nodes. Thus, here comes an n^*n matrix M and M_{ij} is the Euclidean distance between vector $A(V_i)$ and vector $A(V_j)$. In another word, M_{ij} represents the difference between node i and j.

After getting the difference matrix M obtained by the above method, in order to merge the attribute and relationship of node, we define the following operations to transform the difference matrix M into a similarity matrix containing both information (attribute and relationship) of the entire graph. Thus, for each pair of node i and j, $S_{ij}=sim(i, j)$, in which S represents the ultimate similarity matrix. In this experiment, on the base of data density [16], the functions are defined as follows:

$$sim(i,j)=\alpha\times e^{-\frac{W_{ij}}{\sum_{u\in\Gamma(i)}W_{iu}+\sum_{u\in\Gamma(j)}W_{ju}}}+(1-\alpha)\times e^{-\frac{1}{M_{ij}+1}} \tag{8}$$

where $\Gamma(i)$ means the set of adjacent node of i, and α means the similarity coefficient which is usually set as 0.4.

The work of merging the attribute and relationship of node into a similarity matrix in social networks has been completed. By far, an improved preparation work for spectral clustering and subsequent work is finished.

When analyzing social network data with SCBSP, firstly, apply the improved similarity matrix and obtain the similarity matrix S. Secondly, calculate diagonal matrix D using formula (2) and Laplace matrix L of G using formula (3). Thirdly, according to the spectral clustering framework, assuming that the social network is divided into k communities, calculate the largest k eigenvalues and corresponding eigenvectors. Using the k eigenvectors, they form an n^*k matrix, in which the ith row $R_i=(R_{i1}, R_{i2},\ldots,R_{ik})$ represents the information of node i. Finally, apply PSO and SA to obtain the global optimal solution by clustering nodes in the reconstructed sample space using the following method.

It is assumed that the number of particles in particle swarm is m, in which each particle represents a possible community division. The number of dimensions of the particle is the number of nodes in social network. And each dimension of it represents which community the node belongs to. In another word, considering the position vector of ith particle is $X_i=(X_{i1}, X_{i2} \ldots\ldots X_{in})$, each dimension satisfies $1\le X_{ij}\le k$, and X_{ij} means that jth node belongs to X_{ij}th community.

Then, when running the PSO algorithm described above, a little change is made using the strategy of SA. The process updating the global optimal solution P_g has been adjusted because the operation will make the answer fall into a local optima easily just according to the individual optimal solution P_i. So the SA is chosen to approach the global optimal solution P_g after the initial updating operation. In this part, the initial temperature T_{max} in SA is J_c of P_g (J_c is described in Eq. (4)). Here, J_{c_old} represents J_c of P_g. Meanwhile, the reciprocal of initial temperature is set as the minimum temperature, and the annealing coefficient DR and the maximum number of times is set for the inner loop. In the iterations of SA in SCBSP, when one or several cluster results changed which leading to a new division results, the new sum of difference can be gained within cluster J_{c_new}. Then, compare J_{c_now} and J_{c_old}. If J_{c_new} is less than J_{c_old}, then the new dividing results are saved, otherwise calculate the probability p.

$$p=e^{\left(\frac{J_{c\_new}-J_{c\_old}}{a*T_{now}}\right)} \tag{9}$$

where T_{now} is the current temperature and a is a constant. When generating a number between *[0, 1]* uniformly distributed random value r and $r \le p$, then the new state is accepted. Otherwise, the new state is not accepted, and the algorithm continues the iteration. After accepting the new state or if the total looping times in the inner loop reaches the maximum number that we set for the inner loop or the current temperature is lower than the minimum temperature, the process of simulated annealing would be terminated. But if the termination condition has not been satisfied, cool down the temperature using the formula $T_{now}=T_{now}$ * DR, and go on the iteration of the simulated annealing. In the improved version of simulated annealing in SCBSP, there is a certain probability of jumping out current local optimal solution and achieve global optimal solution.

In brief, after merging attribute and relationship of a graph into a single similarity matrix, spectral clustering with PSO and SA is employed to cluster this graph.

4. Experimental study

In order to verify the accuracy of the proposed algorithm, we compared the results of several clustering algorithms such as k-means, Spectral clustering, PSO, Spectral Clustering Based on PSO (SCP) and SCBSP. And the data sources are from Sina microblog, Facebook, as well as randomly generated data. In the preprocessing of data from real-world social networks such as Sina Weibo and Facebook, some isolated points in data collection set are removed in order to ensure good experimental results.

In all of the following experiments, J_c shown in formula (4) is the criteria of evaluating the result of clustering. Apparently, the lower the J_c is, the better the performance a clustering algorithm makes.

Through practice, PSO algorithm performs well in low-dimensional sample space. Therefore, after we compressed information and reduced dimension using spectral clustering, and assisted with the superior ability of SA to find global optimal solution, SCBSP makes excellent performance in the following experiments.

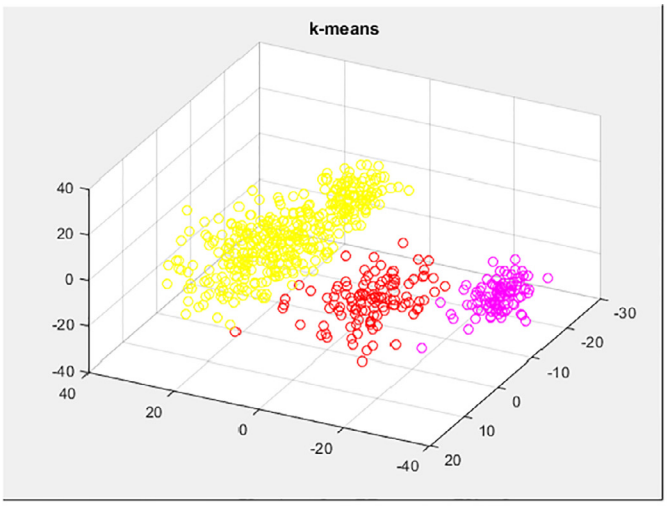
4.1. Results on randomly generated data

In the first experiment, for sake of experimental visibility, we generate three-dimensional random data to verify the performance of SCBSP. In addition, in order to make the comparison between SCP and SCBSP more meaningful and to highlight the superior ability to obtain global optima of SA used in SCBSP, SCP and SCPSP in this experiment both start from the same initial particle swarm.

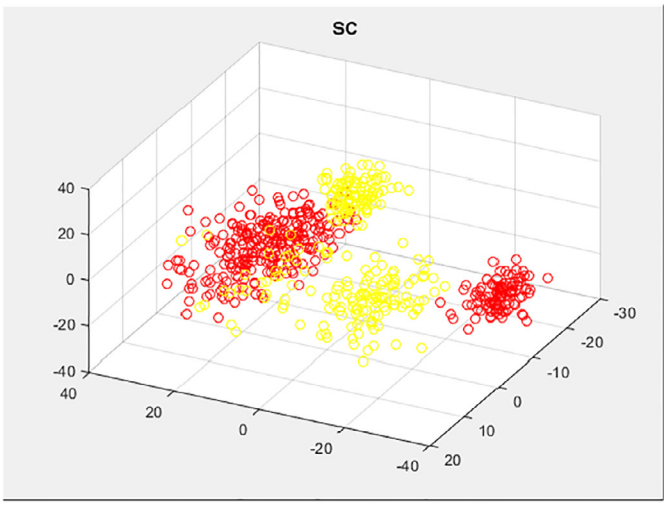
Here, the random data are generated by randomly selecting 6 points at the beginning, then for each point, generating another 100 points according to normal distribution. We choose 4 as the number of clusters because in each small graph in the Fig. 1, human being can identify generally 4 large dense groups of points. Then we run the five different clustering algorithms.

As we can see from Fig. 1 shown above, SCBSP performs the best result among all five different clustering algorithms visually. It is the only algorithm that gets actual 4 clusters, which is closest to the result seen by human being. And in Fig. 2 which compares J_c result and running time of them, SCBSP has both the minimal J_c result and an acceptable running time result.

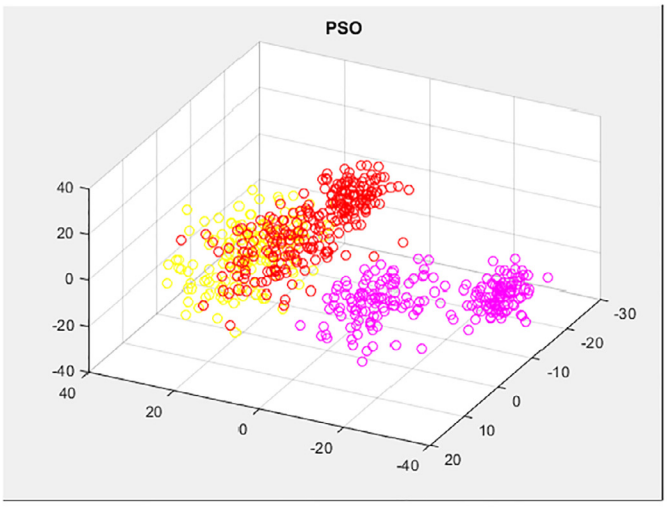
Then, in the second experiment, 30-dimensional random data are used for the comparison among clustering algorithms, with the similar generating method as experiment one, only different in the number of dimensions of points. What's more, relatively higher dimensional data can highlight the good performance of spectral-clustering-based algorithms due to the dimensionality reduction process performed in SC.



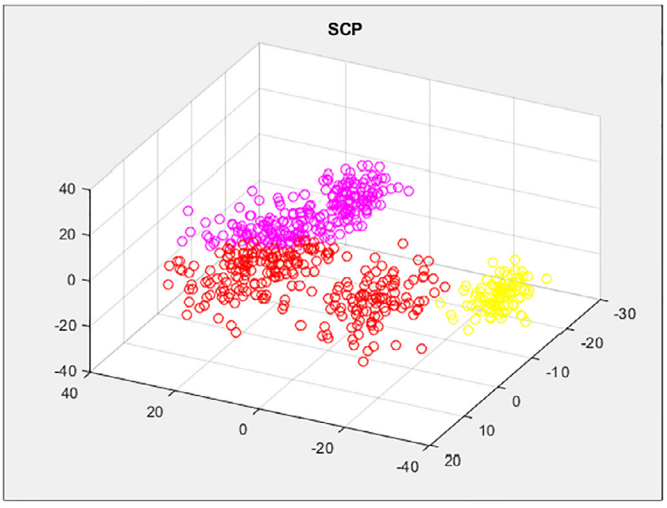
(a) k-means



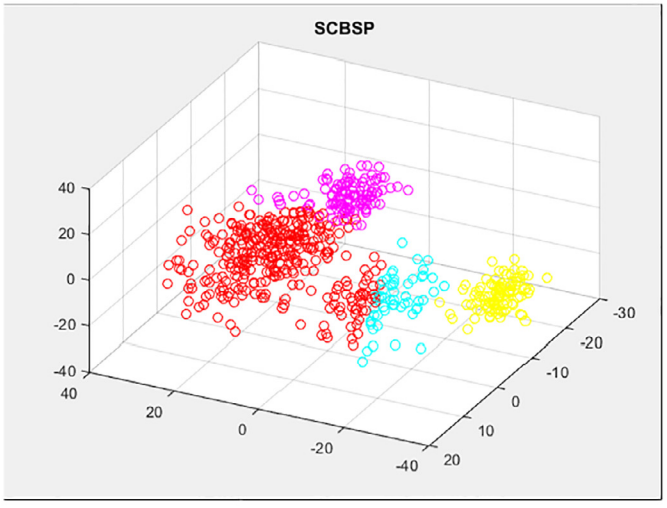
(b) SC



(c) PSO



(d) SCP



(e) SCBSP

Fig. 1. Clustering results of algorithms on 606 three-dimensional points.

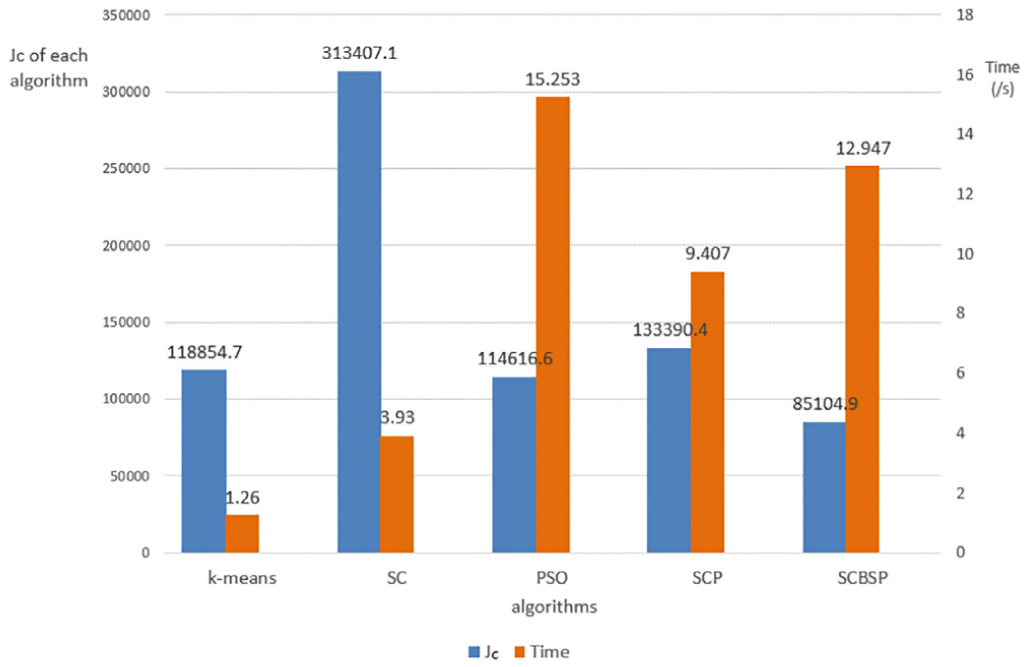


Fig. 2. Clustering result (J_c and Time) of k-means, SC, PSO, SCP and SCBSP.

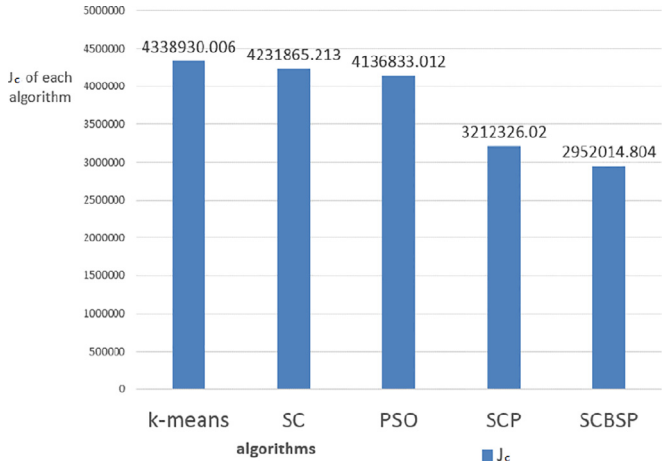


Fig. 3. clustering result (average J_c) of k-means, SC, PSO, SCP and SCBSP on 200 groups of different 606 30-dimensional random points.

For more stable result, we conducted 200 separate experiments and calculated average J_c, and each experiment has 606 different 30-dimensional random points, using the similar data generating strategy employed in experiment one. Then, the average J_c of each algorithm in the 200 experiments is shown in Fig. 3.

In Fig. 3, k-means, original SC, original PSO have similar results, but the other two SC-based algorithms, SCP and SCBSP, perform better apparently. In addition, in the 200 experiments, SCP achieves the minimal J_c among the five algorithms for 69 times, while SCBSP for 108 times, and the J_c of SCBSP is smaller than or equal to that of SCP for 143 times. Therefore, SCBSP performs the best among the five algorithms in this experiment.

In the second experiment described above, clustering performance of the five algorithms have been compared on 30-dimensional random data. Furthermore, we would like to explore the relationship between number of dimensions and the performance of clustering algorithms. So, we make several separate experiments using random points whose number of dimension increases from 2 to 100, and 20 different experiments for each dimension to get an average J_c of each of the five algorithms.

For a better visual experience, the J_c of each algorithm is divided by the J_c of k-means in each experiment of a certain dimension. The result is in Fig. 4.

It is shown in Fig. 4 that SC-based algorithms perform better in higher dimensional sample space, and SCBSP is the best in most dimensions.

Through the experiment above, we can initially verify that the Improved spectral clustering is correct. But the completely verification need the operation on real social network data as follows.

4.2. Results on datasets from Sina microblog and Facebook

Because the nodes in social network themselves do not contain the information about which community it belongs to, the accuracy of the community division cannot be clearly judged. However, the results of the cluster can still be evaluated by the similarity of the nodes within the community, which means the nodes in the same community have the similar properties. In this experiment, the quality of community division by SCBSP is tested with the dataset from the Sina microblog acquired by API. In the network, the information of each user is regard as the attribute of node. Meanwhile, the edge of the graph means the friendship between users. It is assumed that there are 4 clusters, in which condition both SC and SCBSP have best performances. Comparing the performance of SC and SCBSP shown in Fig. 5, SCBSP does better than SC.

In addition to the experiment on dataset from Sina Weibo, we also conducted experiment on a huge dataset from Facebook which contains 4038 nodes and 88,234 edges (Tables 1 and 3).

The strategy of data preprocessing is similar to that in the previous experiment. The only difference is that we introduced “weight” while taking Facebook users’ attribute into consideration. For example, assuming that the full mark is 10, then the weight of “school” is fairly larger number like 8 while the weight of “gender” is relatively smaller number like 3, in that two people who study in the same school is more likely to be friends than two people

Table 1 Algorithm of traditional spectral clustering.

Algorithm: Spectral clustering
Input: 1. A graph $G=(V, E)$ contains information of networks
2. An attribute matrix for n node in G.
Output: Clustering result.
Phase 1(preprocess)
1) Calculate similarity matrix S using formula (1)
2) Calculate diagonal matrix D using the following equation:

$$\begin{cases} D_{ii} = \sum_{j=1}^{n} S_{ji} \\ D_{ij} = 0, i \neq j \end{cases} \tag{2}$$

3) Calculate the Laplace matrix L of G using the following equation:

$$L = D^{-1/2} * S * D^{-1/2} \tag{3}$$

Phase 2
1) Find the largest k eigenvalues and find the corresponding eigenvectors of Laplace matrix L, then construct a new $n*k$ matrix. The jth column of the eigenvectors-matrix is the eigenvector matching the jth largest eigenvalues, and ith row of the matrix is the information of ith node after dimensionality reduction.
2) Cluster with the eigenvector-matrix by k-means clustering.

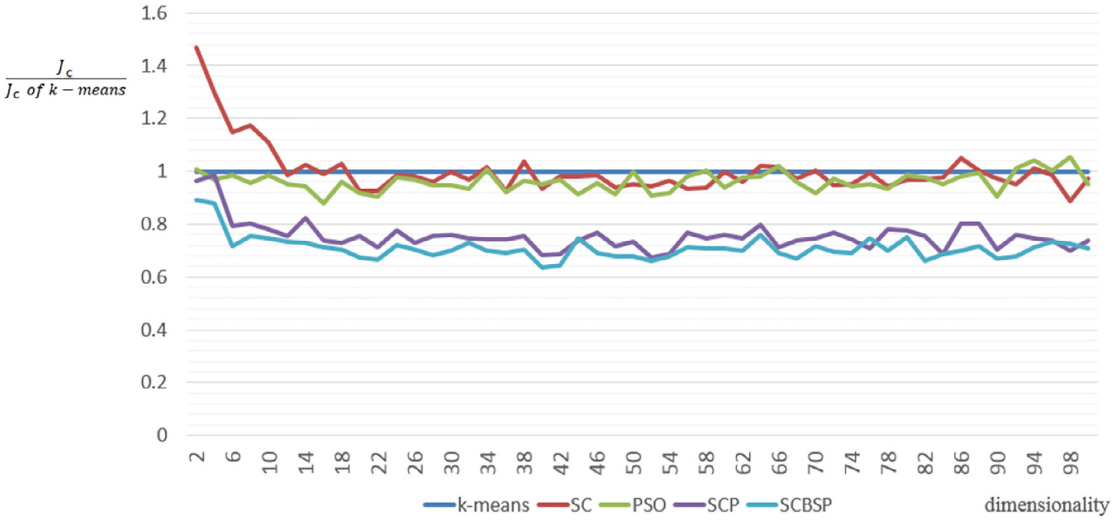


Fig. 4. clustering result (J_c/J_c of k-means) of k-means, SC, PSO, SCP and SCBSP on 606 random points with from 2 dimensions to 100 dimensions.

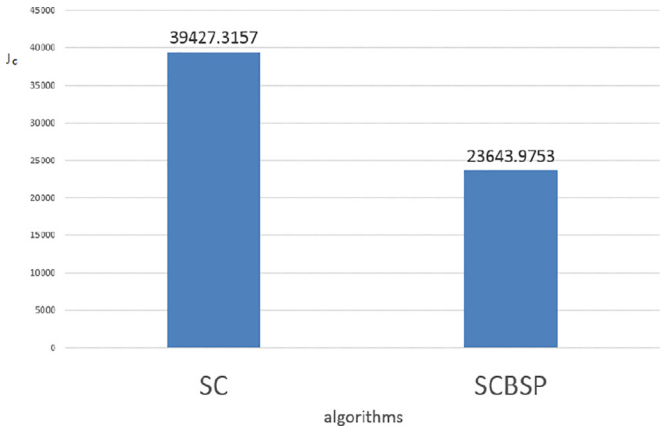


Fig. 5. Clustering result (J_c) of SC, SCP and SCBSP on dataset from Sina Weibo.

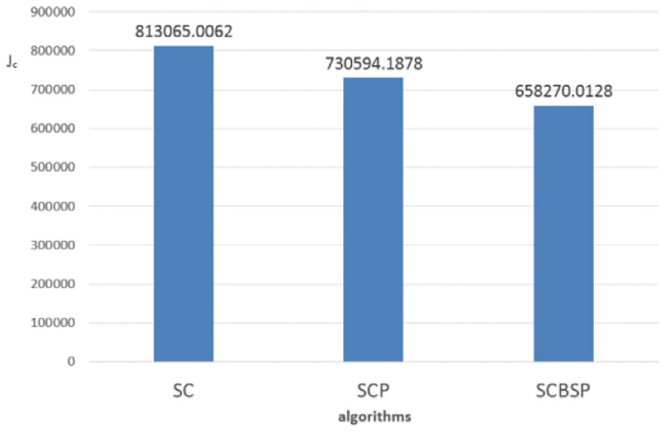


Fig. 6. Clustering result (J_c) of SC, SCP and SCBSP on dataset from Facebook.

who have the same gender. The table of attribute and weight is displayed in Table 2.

Taking both attribute and relationship into consideration, SC, SCP and SCBSP are performed on the dataset and then obtain the following result in Fig. 6, which shows that SCBSP also performs better than SC and SCP on dataset from Facebook.

In the experiment, the relationship between friends is applied in the SCBSP, while the original spectral clustering cannot make full use of this part of information. Taking these factors into consideration, we can conclude that SCBSP has a gigantic advantage over original spectral clustering on community division. From another aspect, the results also show that relationship have a great impact on community division. Taking both attribute and

Table 2 Algorithm of SCBSP.

Algorithm: Spectral Clustering Based on Simulated Annealing and Particle Swarm Optimization
Input: 1. A graph $G=(V, E)$ contains information of social networks
2. An attribute matrix for n node in G.
Output: Community division result.
Phase 1
1) Construct the similarity matrix S by formula (8), diagonal matrix D by formula (2).
2) Calculate Laplace matrix L by formula (3).
3) Find the largest k eigenvalues and corresponding eigenvector R_1, R_2,…,R_k and get an eigenvector matrix.
Phase 2
Initialize the particle swarm and set the maximum steps of iteration, along with the initial random position and random velocity to categorize sample.
Phase 3
1) Calculate the J_c for each particle and update the corresponding individual optimal value P_i.
2) Update the global optimal value P_g with m individual optimal value.
3) Optimize the global optimal value by simulated annealing algorithm.
if iteration times of PSO < maximum number of iterations
i) Update velocity and position of particle by formula (6), (7)
ii) Recalculate the cluster centers
iii) Update the division that which community the nodes belong to
iv) Reapply Phase 3
end if

Table 3
Weight of attributes of users (full mark is 10).

Attribute	Weight	Attribute	Weight
Birthday	3	Location	4
Education; classes	6	Political	7
Concentration	7	Religion	7
Education; degree	6	Work; employer	8
Education; school	8	Work; end_date	5
Education; type	5	Work; from	6
Education; with	5	Work; location	7
Education; year	6	Work; position	6
Gender	3	Work; projects	6
Hometown	5	Work; start_date	5
Languages	4	Work; with	5

relationship of nodes into account can have a great improvement on community division.

5. Conclusion

Since current clustering methods usually do not handle both attribute and relationship well, this paper proposed a new community division algorithm called Spectral Clustering Based on Simulated annealing and Particle swarm optimization (SCBSP) which merges attributes and relationship. The new approach of constructing the similarity matrix in SCBSP not only avoids the use of Radial Basis Function, but also imports the relationship into the similarity matrix. Moreover, SCBSP makes full use of the idea of simulated annealing and particle swarm optimization instead of methods like k-means in the framework of traditional spectral clustering. With the idea of simulated annealing, the proposed algorithm can avoid falling into local optimal solution. Meanwhile, particle swarm optimization provides SCBSP with faster convergence rate. In addition, the facts prove that they increase the effectiveness of SC in community division. From experimental results, merging attribute and relationship of nodes can enhance the accuracy of community division. Compared with the traditional spectral clustering, SCBSP has better application effect in social network.

Though SCBSP has good experimental results, it is based on the fact that a portion of the noise of dataset should be removed before the experiment. If there are more noise nodes of sample collection, the effect of comparison in the experiment will be not ideal enough. This aspect still needs further study. Meanwhile, there still exist some strategies of improvements in the step of simulated annealing. For instance, reducing maximum inner loop times and the maximum number of iteration will lead to different complexity and efficiency. In SCBSP, the number of community division should be known. While in the case of uncertain division number, the aspect still needs further study. In the future, we will also explore more factors that can influent the effectiveness of community division and improve our proposed algorithm. In that merging attribute and relationship of node can greatly affect the clustering result, as we can see from the experiments, there probably exists more factors we have never considered.

Acknowledgments

The research presented in this paper is supported in part by the National Natural Science Foundation of China (Nos. 61650206, 61762002).

References

[1] L. Tang, H. Liu, Community detection and mining in social media, Synth. Lect. Data Min. Knowl. Disco. 2 (1) (2010) 1–137.
[2] A. Ahmad, S. Hashmi, K-Harmonic means type clustering algorithm for mixed datasets, Appl. Soft Comput. 48 (2016) 39–49.
[3] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (6) (2004) 066133.
[4] X. Que, et al., Scalable community detection with the louvain algorithm, in: Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS), IEEE, 2015.
[5] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proceedings of the national academy of sciences 99 (12) (2002) 7821–7826.
[6] C. Shi, et al., A link clustering based overlapping community detection algorithm, Data Knowl. Eng. 87 (2013) 394–404.
[7] S. White, S. Padhraic, A spectral clustering approach to finding communities in graphs, in: Proceedings of the SIAM International Conference On Data Mining, Society for Industrial and Applied Mathematics, 2005.
[8] S. Zhang, R.-S. Wang, X.-S Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Phys. A Stat. Mech. Appl. 374 (1) (2007) 483–490.
[9] J. Shi, et al., Density-based place clustering in geo-social networks, in: Proceedings of the ACM SIGMOD International Conference On Management Of Data, ACM, 2014.
[10] H. Sun, et al., Gskeletonclu: density-based network clustering via structure–connected tree division or agglomeration, in: Proceedings of the IEEE Tenth International Conference on Data Mining (ICDM),, IEEE, 2010.
[11] X. Xu, et al., Scan: a structural clustering algorithm for networks, in: Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2007.
[12] F. Moser, R. Ge, M. Ester, Joint cluster analysis of attribute and relationship data without a-priori specification of the number of clusters, in: Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2007.
[13] D. Verma, M. Meila, A comparison Of Spectral clustering algorithms, University of Washington, 2003, pp. 1–18. Tech. Rep. UWCSE030501 1.
[14] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, NIPS 14 (2) (2001).

[15] C. Aicher, A.Z. Jacobs, A. Clauset, Learning latent block structure in weighted networks, J. Complex Netw. 3 (2) (2014) 221–248.
[16] M.L. Yiu, N. Mamoulis, Clustering objects on a spatial network, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, 2004.
[17] G.-L. Liu, X.-M. Zhen, K-harmonic means clustering with simulated annealing, Comput. Syst. Appl. 7 (2011) 020.
[18] N. Mishra, et al., Clustering social networks, International Workshop on Algorithms and Models for the Web-Graph, Springer, Berlin Heidelberg, 2007.
[19] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[20] R. Kannan, S. Vempala, A. Vetta,, On clusterings: Good, bad and spectral, J. ACM 51 (3) (2004) 497–515.
[21] P. Symeonidis, N. Mantas, Spectral clustering for link prediction in social networks with positive and negative links, Soc. Netw. Anal. Min. 3 (4) (2013) 1433–1447.
[22] Y. van Gennip, et al., Community detection using spectral clustering on sparse geosocial data, SIAM J. Appl. Math. 73 (1) (2013) 67–83.
[23] S. Mehrkanoon, et al., Multiclass semisupervised learning based upon kernel spectral clustering, IEEE Trans. Neural Netw. Learn. Syst. 26 (4) (2015) 720–733.
[24] A. Hadjighasem, et al., Spectral-clustering approach to Lagrangian vortex detection, Phys. Rev. E 93 (6) (2016) 063107.
[25] S. Habashi, N.M. Ghanem, M.A. Ismail, Enhanced community detection in social networks using active spectral clustering, in: Proceedings of the Thirty-First Annual ACM Symposium on Applied Computing, ACM, 2016, pp. 1178–1181.
[26] U. Von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.
[27] F.D. Malliaros, M. Vazirgiannis, Clustering and community detection in directed networks: a survey, Phys. Rep. 533 (4) (2013) 95–142.
[28] H. Jia, S. Ding, M. Du, Self-tuning *p*-spectral clustering based on shared nearest neighbors, Cognit. Comput. 7 (5) (2015) 622–632.
[29] A. LaViers, A.R. Rahmani, M.B. Egerstedt, Dynamic Spectral Clustering, Georgia Institute of Technology, 2010.
[30] E. Aarts, J. Korst, W. Michiels, Simulated annealing, Search Methodologies, Springer, US, 2014, pp. 265–285.

Yi Xu is currently an undergraduate in School of Computer Engineering and Science, Shanghai University. His research interests cover social network analysis and machine learning.

Zhi Zhuang graduated from School of Computer Engineer and Science, Shanghai University and is currently a Computer Science Master in University of California at San Diego. His research interests cover social analysis and machine learning.

Weimin Li is an associate professor in School of Computer Engineering and Science, Shanghai University, China. His research interests cover social network analysis, service computing, and Big data.

Xiaokang Zhou is currently a lecturer of Department of Data Sciences, Shiga University, Japan. He has been engaged in the interdisciplinary research works in the fields of computer science and engineering, information systems, and human informatics.